FIT5195 Major Assignment - SSA/2021-2022 (Weight = 20%)
Due date: Thursday 20-January-2022, 11:55pm

Version: 3.0 – 14//01/2022

Learning Outcomes:

LO1. Design multi-dimensional databases and data warehouses.
LO2. Use fact and dimensional modelling.
LO3. Implement online analytical processing (OLAP) queries.
LO4. Explain the roles of data warehousing architecture and the concepts of granularity in data warehousing.
LO5. Propose business intelligence reports using data warehouses and OLAP.

A. General Information and Submission

- This is an individual assignment.
- *Submission method*: Submission is online through Moodle.
- *Penalty for late submission*: 10% deduction for each day.
- *Assignment FAQ*: There is a Major Assignment Frequently Asked Questions page set up for the Major Assignment on EdStem Forum.

B. Problem Description

M-Stay is a residential service that offers homestay and rental services to Monash students and staff around Melbourne. The company has an existing operational database that maintains and stores all of the business transactions information (e.g. properties, hosts, listings, booking, etc.) required for the management's daily operation. As the business grows, M-Stay has decided to build a Data Warehouse to improve their analysis and work efficiency. However, since the staff at M-Stay have limited Business Intelligence and Data Warehouse knowledge, they have decided to hire you to design, develop and quickly generate BI reports from a Data Warehouse.

The operational database tables can be found at the **MStay** account. You can, for example, execute the following query:

```
select * from MStay.<table_name>;
```

The data definition of each table in MStay is as follows:

Table Name	Attributes and Data Types	Notes

REVIEW	Review_ID	Number	The table stores review information of the related booking order.
	Review_Date	Date	
	Review_Comment	Varchar	
	Booking_ID	Number	
BOOKING	Booking_ID	Number	The table stores booking information.
	Booking_Date	Date	
	Booking_Stay_Start_Date	Date	
	Booking_Duration	Number	
	Booking_Cost	Number	
	Booking_Num_Guests	Number	
	Listing_ID	Number	
	Guest_ID	Number	
GUEST	Guest_ID	Number	The table stores all guest information.
	Guest_Name	Varchar	
LISTING	Listing_ID	Number	The table stores all listing information. Each listing has one property and one host information.
	Listing_Date	Date	
	Listing_Title	Varchar	
	Listing_Price	Number	
	Listing_Min_Nights	Number	
	Listing_Max_Nights	Number	
	Prop_ID	Number	

	Type_ID	Number	
	Host_ID	Number	
HOST	Host_ID	Number	The table stores all host information.
	Host_Name	Varchar	
	Host_Since	Date	
	Host_Location	Varchar	
	Host_About	Varchar	
	Host_Listing_Count	Number	
HOST_VERIFICATION	Host_ID	Number	The table stores the verification information between host and channel.
	Channel_ID	Number	
CHANNEL	Channel_ID	Number	The table stores the channel of verification for the hosts.
	Channel_Name	Varchar	
LISTING_TYPE	Type_ID	Number	The table stores all listing types.
	Type_Description	Varchar	
PROPERTY	Prop_ID	Number	The table stores all property information.
	Prop_Description	Varchar	
	Prop_Neighbourhood_Overview	Varchar	
	Prop_Num_Beds	Number	
	Prop_Num_Bedrooms	Number	
	Prop_Num_Bathrooms	Number	
	Prop_Num_Reviews	Number	

	Prop_Rating_Location	Number	
	Prop_Rating_Cleanliness	Number	
	Prop_Rating_Value	Number	
	Prop_Average_Rating	Number	
PROPERTY_AMMENITY	Prop_ID	Number	The table links property and amenity tables
	Amm_ID	Number	
AMENITY	Amm_ID	Number	The table stores all amenities information
	Amm_Description	Varchar	

C. Tasks

The assignment is divided into **FOUR** main tasks:

1. **Design a data warehouse for the above M-Stay database.**
   You are required to create a data warehouse for the **M-Stay** database.
   The management is especially interested in the following fact measures:

   - Number of properties
   - Number of hosts
   - Average listing price

   The following show some possible dimension attributes that you should need in your data warehouse:

   - Listing type
   - Listing time [Month, Year]
   - Listing season
   - Listing maximum stay duration [*short*: less than 14 nights, *medium*: 14 to 30 nights, *long*: more than 30 nights]

- Property size based on the number of beds and bedrooms [*small*: minimum of 1 bed and 1 bedroom, *medium*: minimum of 3 beds and 2 bedrooms, *large*: more than 5 beds and more than 3 bedrooms]
- Property rating
- Host channels
- Host join time [Month, Year]
- Host location

For each attribute, you may apply your own design decisions on specifying a range or a group, but make sure to specify them in your submission.

- **Preparation stage.**

  Before you start designing the data warehouse, you have to ensure that you have explored the operational database and have done sufficient data cleaning. Once you have done the data cleaning process, you are required to explain what strategies you have taken to explore and clean the data.

  The outputs of this task are:

  a) The E/R diagram of the operational database,

  b) If you have done the data cleaning process, explain the strategies you used in this process (you need to show the SQL to explore the operational database and SQL of the data cleaning, as well as the screenshot of data *before* and *after* data cleaning),

- **Designing the data warehouse by drawing star/snowflake schema.**

  The star schema for this data warehouse contains multi-facts. You need to identify the fact measures, dimensions, and attributes of the star/snowflake schema. The following queries might help you to determine the fact measures and dimensions:

  - How much is the total listing price for “Private room” in Summer?
  - How many listings are listed in June 2015?
  - How much is the average listing price for short-term stay duration?
  - What is the average listing price of 5-star properties?
  - How much is the total listing price for the listings belonging to the host who are located at Healesville, Victoria, Australia?
  - How many 4-star properties are in medium size?
  - How many hosts are located at South Yarra, Victoria, Australia?
  - How many hosts use email and phone channels?

You should pay attention to the granularity of your fact tables. You are required to create **two versions** of star/snowflake based on different levels of aggregation.
The two versions of the star/snowflake represent different levels of aggregation. Version-1 should be at the highest level of aggregation. Version-2 should be in level 0, which means no aggregation. To make it simple, you can assume that the highest aggregation for this assignment is Level-2.

Version Name	Level
Version-1	High aggregation (Level 2)
Version-2	No aggregation (Level 0)

The star/snowflake schema of both versions you created might contain **Bridge Table** and **Temporal**. If a bridge table is needed, you will need to include GroupList and WeightFactor attributes in the relevant dimension. If a temporal dimension is needed, you can use any suitable temporal data warehousing techniques for the temporal dimension and provide the reasons for your choice.

The outputs of this task are:

c) Two versions of star/snowflake schema diagrams,

d) The reasons for the choice of SCD type for temporal dimension, if any,

e) A short explanation of the difference between the two versions of the star/snowflake schema.

2. **Implement version 1 star/snowflake schema using SQL.**
You are required to implement the star/snowflake schema for version 1 that you have drawn in Task 1. This implies that you need to create the fact and dimension tables for version 1 in SQL. The output is a series of SQL statements to perform this task. You will also need to show that this task has been carried out successfully.

Note:

- If your account is full, you will need to drop all of the tables that you have previously created during the tutorials.
- If you have dopped all tables in you account and you still encounter the `ORA-01536: space quota exceeded for tablesace 'TABLE_NAME'`, please check your SQL code whether you have properly joined all tables. This issue was mainly caused when you did not do the table join properly as the number of records multiplied during the process.

The outputs of this task are:

a) SQL statements (e.g. create table, insert into, etc) to create the star/snowflake schema Version-1.

b) Screenshots of the tables you have created; this includes the contents of each table that you have created. If the table is very big, you can show only the first part of the data.

3. **Create the following reports using OLAP queries.**
You are required to generate the reports using data warehouse **version-1 (Level 2)** that you have implemented in Task 2. For each report, you ought to produce the SQL command and sample report output.

   a. *Simple reports:*

   Produce **two** reports. Each report contains two attributes from two different dimensions and one fact measurement.

   For the report itself, the first report must be about **Top *n*** and the second report is **Top *n%*.**

   The outputs of this task are:

   (a) The query questions that are written in English,
   (b) Your explanation on why such a query is necessary or valuable for the management,
   (c) The SQL commands, and
   (d) The screenshots of the query results (or part of the query results), including all attribute names.

   b. *Reports with proper sub-totals:*

   Produce **four** reports. These reports must include subtotals, using the Cube or Roll-up or Partial Cube/Roll-up operators.

   REPORT 3 and REPORT 4: What are the subtotals and total listing price from each listing type, season, and listing duration? (You must use the Cube and Partial Cube operator)

   REPORT 5 and REPORT 6: Produce two other subtotals reports that are useful for management using Roll-up and Partial Roll-up

   The outputs of this task are:

   (a) The query questions that are written in English,
   (b) Your explanation on why such a query is necessary or valuable for the management,

(c) The SQL commands that include subtotals, using the Cube or Roll-up or Partial Cube/Roll-up operators, and
(d) The screenshots of the query results (or part of the query results).

c. *Reports with moving and cumulative aggregates:*

Produce **two** reports containing moving and cumulative aggregates.

REPORT 7: What are the total listing price and cumulative total listing price of small properties in each year?

REPORT 8: Produce another moving/cumulative aggregate report that is useful for management.

The outputs of this task are:

(a) The query questions that are written in English,
(b) Your explanation on why such a query is necessary or valuable for the management,
(c) The SQL commands that contain moving and cumulative aggregates, and
(d) The screenshots of the query results (or part of the query results).

d. *Reports with Partitions:*

Produce **two** reports that contain partitions.

REPORT 9: Show ranking of each property size and ranking of each property rating based on the total number of properties.

REPORT 10: Produce another partitioning report that is useful for management.

The outputs of this task are:

(a) The query questions that are written in English,
(b) Your explanation on why such a query is necessary or useful for the management,
(c) The SQL commands that contain partitions, and
(d) The screenshots of the query results (or part of the query results), including all attribute names.

4. **Business Intelligence (BI) Reports.**
Choose at least **three** reports from Task 3, and change the presentation of these reports by representing these in a BI report format. Create **one** dashboard based on your chosen reports. This new presentation should be more appealing to the

management. You can use any visualisation tools (e.g. Oracle Report, PowerBI, Tableau) to show the BI reports. Additionally, in these new reports, you might want to include some selection buttons (for illustrative purposes), which may give users options on what criteria to choose so that the graph report will be more dynamic.

D. Checkpoints

There will be checkpoints in Session 6 and 8:

Checkpoint	Weight	Assessment	Deadline
Checkpoint 1	1%	ER Diagram Data Cleaning	Session 6 (during lab or by appointment)
Checkpoint 2	1%	Star Schema v1	Session 8 (during lab or by appointment)

You are required to complete the assessment for a given checkpoint in order to obtain the allocated mark. There are associated mark penalties for not meeting the checkpoint assessment on time to a satisfactory state.

If you are unable to attend the allocated checkpoint, please contact your tutor immediately.

Note that the Final Report and Code are worth 18%.

E. Submission Checklist

1. One **combined pdf file** containing all tasks mentioned above:
   - □ Cover page
   - □ Task C.1 (outputs a, b, c, d, e)
   - □ Task C.2 (outputs a, b)
   - □ Task C.3 Simple Reports (outputs a, b, c, d)
   - □ Task C.3 Reports with Subtotals (outputs a, b, c, d)
   - □ Task C.3 Reports with Moving and Cumulative Aggregates (outputs a, b, c, d)

- □ Task C.3 Reports with Partitions (outputs a, b, c, d)
- □ Task C.4 (a dashboard with at least three reports)

2. **.sql files** for the following task:
    - □ Task C.1 (SQL command as required by output *b*)
    - □ Task C.2 Implement Star Schemas (SQL command as required by output a)
    - □ Task C.3 Simple Reports (SQL command as required by output c)
    - □ Task C.3 Reports with Subtotals (SQL command as required by output c)
    - □ Task C.3 Reports with Moving and Cumulative Aggregates (SQL command as required by output c)
    - □ Task C.3 Reports with Partitions (SQL command as required by output c)

    All of the above SQL files must be runnable in Oracle.

3. The file of the BI tool of your choice for the following task:
    - □ Task C.4 (a dashboard with at least three reports)

4. Zip all the SQL files from #2 and the file from #3, and name the ZIP folder as **MA_SQL_BI.zip**.

Submission Method:

1. Upload the **PDF file** from Checklist #1 and the **ZIP file** from Checklist #4 to Moodle by the due date: **Thursday, 20 January 2022, 11:55pm**.
    - The submission of this assignment must be in the form of **a single PDF file AND a single ZIP file**. No other forms will be accepted.
    - You must ensure that you have all the files listed in this checklist before submitting your assignment to Moodle. Failure to submit a complete list of files will lead to mark penalties.
2. Penalty for late submission: 10% deduction for each day, including weekends.
3. Submission Cut-off time: **Thursday, 27 January 2022, 11:55 pm** (Submission link will be unavailable after the cut-off date).

Getting help and support:

What can you get help for?

- ***Consultations with the Teaching Team***
  Talk to the Teaching Team:
  https://lms.monash.edu/course/view.php?id=132353&section=2
- ***English language skills***
  Talk to English Connect: https://www.monash.edu/english-connect
- ***Study skills***
  Talk to a learning skills advisor: https://www.monash.edu/library/skills/contacts
- ***Counselling***
  Talk to a counsellor: https://www.monash.edu/health/counselling/appointments

Extensions:

If you are experiencing difficulties that you think will impact your ability to meet this deadline, you may apply for an assignment extension. You must apply **no later than two University working days after the due date** of this assignment.

The extension application can be found on *Moodle > Assessments > How to Apply for an Extension*. Please allow **two business days** for your application to be processed.

Please ensure your application is supported by appropriate documentation. You can find more information about assignment extensions at the Special Consideration website.

Special Considerations:

Students should carefully read the Special Consideration website, especially the details about what formal documentation is required.

All special consideration requests should be made using the Special Consideration Application.

Please do not assume that submission of a Special Consideration application guarantees that it will be granted – you must receive an official confirmation that it has been granted.

Late Penalty:

Late assignments submitted without an approved extension may be accepted (up to a maximum of **seven days**) with the approval of the Chief Examiner and/or Lecturer but will be **penalised at the rate of 10% per day (including weekends and public holidays)**. Assignments submitted more than seven days after the due date will receive a zero mark for that assignment and may **not receive any feedback**.

Plagiarism and Collusion:

Monash University is committed to upholding standards and academic integrity and honesty. Please take the time to view these links.

Academic Integrity Module

Student Academic Integrity Policy

Test your knowledge, collusion (FIT No Collusion Module)

END OF MAJOR ASSIGNMENT